

# Fake News Detection System

Rajendra Chatse<sup>1</sup>, Pradeepkumar Kale<sup>2</sup>, Nikhil Choudhari<sup>3</sup>, Adesh Shinare<sup>4</sup>, Sarthak Kale<sup>5</sup>

<sup>1</sup>Professor, Department of Information Technology, SVPM College of Engineering, Baramati, India

<sup>2,3,4,5</sup>Student, Department of Information Technology, SVPM College of Engineering, Baramati, India

**Abstract:** The things you read online especially on social media may occur to be true or fake. Fake news is a news, stories or hoaxes that are created purposely to misinform or betray the readers. So, we design this paper for a feature of detecting a fake news by using different machine learning techniques. In this domain of machine learning algorithm, we use n-gram analysis for fake news detection. This strategy utilizes NLP Classification model to anticipate whether a post on Twitter will be named as REAL or FAKE. We propose in this system, a phony news model that utilization naive bayes algorithm.

**Keywords:** NLP, Text Classification, Naive Bayes.

## 1. Introduction

In the ongoing years, online substance has been assuming a huge job in influencing client's choices and suppositions. Counterfeit news is a marvel which is significantly affecting our public activity, specifically in the political world. Counterfeit news location is a rising exploration region which is picking up intrigue yet included a few difficulties because of the restricted measure of assets available. Information accuracy on Internet, particularly via web-based networking media, is an undeniably significant concern, however web-scale information hampers, capacity to distinguish, assess and right such information, or supposed "counterfeit news," present in these stages. In this paper, we have exhibited a recognition model for phony news utilizing NLP investigation through the Sentiment Analysis strategies. The proposed model accomplishes its most elevated exactness. Counterfeit news discovery is a developing exploration region with couple of open datasets. In our model we develop a system for check whether the news is counterfeit or not on twitter by using NLP and Naive bayes. Multiple news channels dataset available to check counterfeit news.

## 2. Aim and objective

Detect the various news on twitter fake or real prediction. The main goal of this project is to design Detection of Online Fake News Using NLP (Natural Language Processing) and Machine Learning Technique. The project is concerned with identifying a solution that could be used to detect and filter out sites containing fake news.

## 3. Literature survey

*Reference No: 1*

*Title:* "Detection of Online Fake News Using N-Gram

Analysis and Machine Learning Techniques"

*Author:* Hadeer Ahmed (2017)

*Publisher:* Springer

*Summary:* Detecting fake news is believed to be a complex task given that humans tend to believe misleading information and the lack of control of the spread of fake content. In this paper SVM algorithm can be chosen from different machine learning techniques with high accuracy 92 percent.

*Reference No: 2*

*Title:* Detecting Fake News in Social Media Networks.

*Authors:* Monther Aldwairi, Ali Alwahedi

*Publisher:* ScienceDirect

*Summary:* The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts. The experimental results show a 99.4 percent accuracy using logistic classier.

Companies such as Google, Facebook, and Twitter have attempted to address this particular concern. However, these efforts have hardly contributed towards solving the problem as the organizations have resorted to denying the individuals associated with such sites the revenue that they would have realized from the increased track. Users, on the other hand, continue to deal with sites containing false information and whose involvement tends to act the reader's ability to engage with actual news. The reason behind the involvement of rms such as Facebook in the issue concerning fake news is because the emergence and subsequent development of social media platforms have served to exacerbate the problem.

*Reference No: 3*

*Title:* Fake News Detection

*Authors:* Akshay Jain, Amey Kasbe

*Publisher:* IEEE

*Summary:* This paper describes a simple fake news detection method based on one of the machine learning algorithms – naive Bayes classifier. The goal of the research is to examine how naive Bayes works for this particular problem, given a manually labelled news dataset, and to support (or not) the idea of using artificial intelligence for fake news detection. Further, this technique can easily be applied to social platforms like Facebook and Twitter by adding recent news and enhancing the

dataset on a regular basis. The difference between this paper and other papers on the similar topics is that in this composition naive Bayes classifier was specifically used for fake news detection; we have tested the difference in accuracy by taking different length of articles for detecting the fake news; also a concept of web scrapping was introduced which gave us an insight into how we can update our dataset on regular basis to check the truthfulness of the recently updated Facebook posts.

#### 4. Mathematical model

Let S be the Whole system which consists:

➤  $S = \{\text{News}\}$

Where,

News is the input of the system.

➤ *Input:*

$IP = \{I\}$

Where,

I is set of News provided as an input.

➤ *Procedure:*

- Step 1: News are retrieved the dataset into the Twitter.
- Step 2: Verify the information into database.
- Step 3: We build up a straightforward NLP based classier to separate among phony and genuine news stories.
- Step 4: Show result.

➤ *Output:*

News is fake or real.

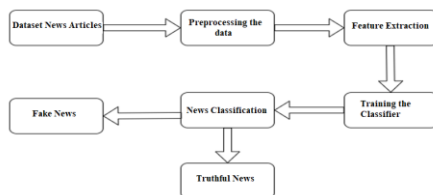


Fig. 1. System architecture

#### 5. Algorithm

##### 1) NLP

*Step 1: Sentence Segmentation*

Breaking the piece of text in various sentences.

*Step 2: Word Tokenization*

Breaking the sentence into individual words known because the tokens. We have a tendency to tokenize them whenever we have encounter a space, we have train a model therein approach. Even punctuations are considered through about as individual tokens as they are have some meaning.

*Step 3: Predicting Parts of Speech for each token*

Predicting whether the word is a noun, verb, adjective, adverb, pronoun, etc. This may be help to understand what the sentence is talking regarding it. This could be achieved by feeding the tokens (and the words around it) to a pre-trained part-of speech classification model. This model was fed plenty a lot of English words with various components of speech

labelled to them so that it classifies the similar words it encounters in future in various components of speech. Again, the models don't really understand the 'sense' of the words, it just simply classifies them on the idea of its previous expertise. It's pure statistics.

*Step 4: Lemmatization*

Feeding the model with the basis root word.

*Step 5: Identifying stop words*

There are numerous words in the English language that are measure used often like 'a', 'and', 'the' etc. These words build plenty a lot of noise whereas doing statistical analysis. We are able take these words out. Some language processing pipelines can categorize these words as stop words, they are going be filtered out while doing some statistical analysis. Definitely, they are needed to understand the dependency between various tokens to get the exact sense of the sentence. The list of stop words varies and depends on what reasonably (kind of) output are you expecting.

*Step 6.1: Dependency Parsing*

This means finding out link or relationship between the words within the sentence and how they are associated with one another. We have a tendency to produce parse tree in dependency parsing, with root because the main verb within the sentence. If we consider talk about tendency the first sentence in our example, then 'is' is that the main verb and it have been the root of the parse tree. We can able to construct a parse tree of each sentence with one root word (main verb) related to it. We also able to find out kind of relationship that exists between the 2 words. In our example, 'San Pedro' is that the subject and 'island' is that the attribute. Thus, the relationship between 'San Pedro' and 'is', and 'island' and 'is' will be established.

*Step 6.2: Finding Noun Phrases*

We are able to cluster the words that represent the identical plan. For eg. It's the second-largest city within the Belize District and largest within the Belize Rural South constituency. Here, tokens 'second', 'largest' and 'town' will be grouped together as they together represent the identical(same) thing 'Belize'. We are able to use the output of dependency parsing to mix such words. Whether to try this step or not fully depends on the end goal, but it's always quick to do this if don't want information about which words are adjective, rather concentrate on other important details.

##### 2) Naive Bayes

Naïve baye's particularly helpful for very large massive data sets. Together with simplicity, Naive Bayes is known to even high sophisticated extremely classification ways.

Bayes theorem provides the way of scheming (calculating) posterior probability  $P(c|x)$  from  $P(c)$ ,  $P(x)$  and  $P(x|c)$ . Check out the equation below:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood
Class Prior Probability

Posterior Probability
Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

$P(c|x)$  is that the posterior probability of the class (c, target) given predictor (x, attributes).

$P(c)$  is that prior probability of the class.

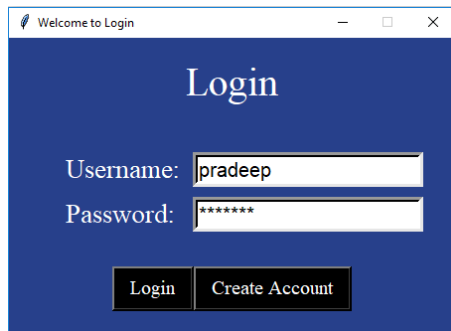
$P(x|c)$  is that the chance that is likelihood which is the probability of predictor given category class.

$P(x)$  is that prior likelihood of predictor.

### 6. Working models and results

1. First System Login
2. Then Registration
3. Twitter Data Scraping
4. Twitter data to csv conversion
5. Applying NLP Algorithm and predict the Positive, Negative and Neutral
6. Fake News Detection.

#### 1) Login Page

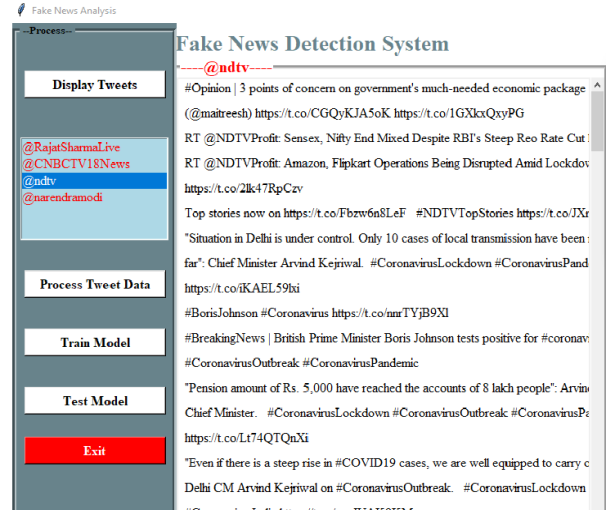


#### 2) Sign-Up Page



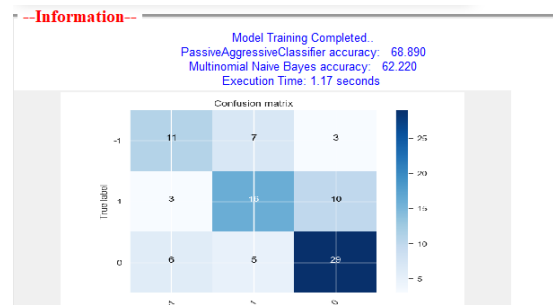
#### 3) Extracting the Tweets

Accessing all the tweets of the particular channel. Process tweet data, after processing train model and then test model.



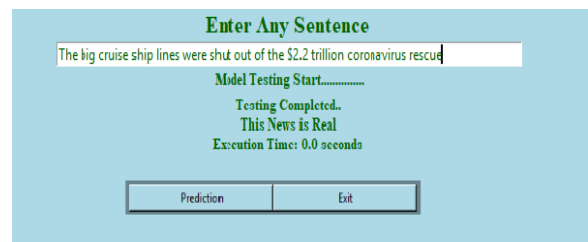
#### 4) Trained model

Showing the confusion matrix and accuracy of the classifier.



#### 5) Test Model

Search any tweet for prediction of Fake or Real and test model.



### 7. Conclusion

In this model, we have introduced model for phony news utilizing Machine Learning techniques and NLP analysis through the Semantic Analysis strategies. The proposed model accomplishes its most elevated exactness. Counterfeit news discovery is a developing exploration zone with couple of

twitter channels and predict the result fake or real.

### References

- [1] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, Kiev, 2017, pp. 900-903.
- [2] A. Gupta and R. Kaushal, "Improving spam detection in Online Social Networks," *2015 International Conference on Cognitive Computing and Information Processing (CCIP)*, Noida, 2015, pp. 1-6.
- [3] Lemann, N, Solving the Problem of Fake News. *The New Yorker* (2017). <http://www.newyorker.com/news/news-desk/solving-the-problem-of-fake-news>
- [4] Essay: The Impact of Social Media. <https://www.ukessays.com/essays/media/the-impact-of-soc>